Erratum to *Incremental Syntactic Language Models for Phrase-based Translation*

Lane Schwartz
Air Force Research Laboratory
Wright-Patterson AFB, OH USA
lane.schwartz@wpafb.af.mil

Chris Callison-Burch
Johns Hopkins University
Baltimore, MD USA
ccb@cs.jhu.edu

William Schuler
Ohio State University
Columbus, OH USA
schuler@ling.ohio-state.edu

Stephen Wu
Mayo Clinic
Rochester, MN USA
wu.stephen@mayo.edu

Abstract

Schwartz et al. (2011) presented a novel technique for incorporating syntactic knowledge into phrase-based machine translation through incremental syntactic parsing, and presented empirical results on a constrained Urdu-English translation task. The work contained an error in the description of the experimental setup, which was discovered subsequent to publication. After correcting the error, no improvement in BLEU score is seen over the baseline when the syntactic language model is used on the constrained Urdu-English translation task. The error does not affect the originally reported perplexity results.

1 Error

Schwartz et al. (2011) presented a novel technique for incorporating syntactic knowledge into phrase-based machine translation through incremental syntactic parsing. That work contained an error in the description of the experimental setup, which was discovered subsequent to publication. The penultimate sentence of Section 6 stated that during MERT (Och, 2003), "we tuned the parameters using a constrained dev set (only sentences with 1-20 words)." While this was the intended experimental configuration, subsequent to publication a re-examination of the experiment revealed that for the condition where the HHMM syntactic language model was used in addition to the n-gram language model (HHMM + n-gram), tuning was actually performed using a constrained dev set of sentences with 1-40 words.

As a result of this error, the BLEU scores reported in Figure 9 do not represent directly comparable experimental conditions, since the dev set used for tuning was different (sentences with 1-20 words for n-gram only versus sentences with 1-40 words for HHMM + n-gram).

Because the results are not comparable, the claims of statistically significant improvements to translation quality are not justified. In order to provide comparable results, we re-ran the n-gram only configuration performing tuning with a constrained dev set of 1-40 words, to match the actual configuration that was used for the HHMM + n-gram configuration. A list of corrections is listed below.

2 List of Corrections

- Abstract, final sentence:

  > We present empirical results on a constrained Urdu-English translation task that demonstrate a significant BLEU score improvement and a large decrease in perplexity.

  should become

Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the sponsors or the United States Air Force. Material presented here was cleared for public release with Case Number 88ABW-2010-6489 on 10 Dec 2010 and with Case Number 88ABW-2012-0302 on 19 Jan 2012.

> We present empirical results on a constrained Urdu-English translation task that demonstrate a large decrease in perplexity but no significant improvement to BLEU score.

- Section 1, final sentence:

  > Integration with Moses (§5) along with empirical results for perplexity and significant translation score improvement on a constrained Urdu-English task (§6)

  should become

  > Integration with Moses (§5) along with empirical results for perplexity and translation scores on a constrained Urdu-English task (§6)

- Section 6, final two sentences:

  > Due to this slowdown, we tuned the parameters using a constrained dev set (only sentences with 1-20 words), and tested using a constrained devtest set (only sentences with 1-20 words). Figure 9 shows a statistically significant improvement to the BLEU score when using the HHMM and the n-gram LMs together on this reduced test set.

  should become

  > Due to this slowdown, we tuned the parameters using a constrained dev set (only sentences with 1-40 words), and tested using a constrained devtest set (only sentences with 1-20 words). Figure 9 shows no statistically significant improvement to the BLEU score when using the HHMM and the n-gram LMs together on this reduced test set.

- Figure 9:

Moses LM(s)	BLEU
n-gram only	18.78
HHMM + n-gram	**19.78**

should become

Moses LM(s)	BLEU
n-gram only	21.43
HHMM + n-gram	21.72

- Section 7, sentence 5:

  > The translation quality significantly improved on a constrained task, and the perplexity improvements suggest that interpolating between n-gram and syntactic LMs may hold promise on larger data sets.

  should become

  > While translation quality did not significantly improve on a constrained task, the perplexity improvements suggest that interpolating between n-gram and syntactic LMs may hold promise on larger data sets.

3 Conclusion

The description of the experimental setup in Schwartz et al. (2011) contained an error that was discovered subsequent to publication. The description stated that MERT was performed on a constrained dev set of sentences with 1-20 words. In fact, one of the experimental conditions (HHMM + n-gram) was instead run on a constrained dev set of sentences with 1-40 words. This error has been corrected — after correction, no statistically significant improvement to translation quality is seen in terms of BLEU score. The error does not affect the originally reported perplexity results.

References

Franz Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July.

Lane Schwartz, Chris Callison-Burch, William Schuler, and Stephen Wu. 2011. Incremental syntactic language models for phrase-based translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 620–631, Portland, Oregon, June.